DOCUMENT RESUME

ED 049 318 TM 000 519

AUTHOR	Nitko, Anthony J.
TITLE	A Model for Criterion-Referenced Tests Based on Use.
INSTITUTION	Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	Feb 71
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Behavioral Objectives, Correlation, *Criterion Referenced Tests, Diagnostic Tests, Individual Characteristics, *Individualized Instruction, *Instructional Design, Item Analysis, Models, Predictor Variables, Scores, *Test Construction, *Tests

ABSTRACT

The nature and purpose of criterion-referenced testing is discussed in light of test design procedures. It is seen that the uses to which test results are put are the chief determiners of the appropriate measurement model. A distinction is made between cut-off scores, criterion scores, and mastery scores. The value of certain test construction procedures in designing criterion-referenced tests for use in adaptive individualized instructional systems is discussed and cautions in the use of traditional procedures are rated. It is concluded that traditional procedures cannot be avoided in some instances, but must be avoided in others. (Author)

ED049318

A Model for Criterion-Referenced Tests Based on Use

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Anthony J. Nitko
Learning Research and Development Center
University of Pittsburgh

TM 000 519

A paper presented at the Annual Meeting of the American Educational Research Association, New York City, New York, February 4-7, 1971.

The preparation of this paper was supported by the Learning Research and Development Center supported as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare.

Abstract

A Model for Criterion-Referenced Tests Based on Use

by

Anthony J. Nitko

University of Pittsburgh

The nature and purpose of criterion-referenced testing is discussed in light of test design procedures. It is seen that the uses to which test results are put are the chief determiners of the appropriate measurement model. A distinction is made between cut-off scores, criterion scores, and mastery scores. The value of certain test construction procedures in designing criterion-referenced tests for use in adaptive individualized instructional systems is discussed and cautions in the use of traditional procedures are rated. It is concluded that traditional procedures cannot be avoided in some instances, but must be avoided in others.

A. J. Nitko
2/71

A Model for Criterion-Referenced Tests Based on Use

When the term "criterion-referenced test" is used (e.g., by Glaser and Klaus, 1962; Glaser, 1963; Glaser and Cox, 1968) it has a somewhat different meaning from the two more prevalent uses of the terms criterion or criterion tests in educational and psychological literature. One of these usages involves the notion of a correlation of scores, X, with a second set of scores, Y. The Y-scores, which may be a second test or performance rating, for example, are often termed criterion scores. The degree to which the X-scores relate to the criterion Y-scores is often expressed by some type of correlation coefficient.

A second interpretation of the term criterion concerns the imposition of an acceptable score magnitude as an index of attainment. Phrases such as "working to criterion-level" and "mastery is indicated by obtaining a score equivalent to 90 per cent of the items correct," are indicative of this type of interpretation of criterion.

Neither of these two types of interpretations is quite what is meant by a criterion-referenced test. A criterion-referenced test is one that is deliberately constructed to yield scores that are directly interpretable in terms of specified performance standards (Glaser and Nitko, in press). Thus, "the standard [or criterion] against which a student's performance is compared . . . is the behavior which defines each point along the achievement continuum (Glaser, 1963, p. 519)." Four things are characteristic of criterion-referenced tests (Nitko, 1970):

(1) the classes of behavior that define different achievement levels are specified as clearly as is possible before the test is constructed.

(2) each behavior class is defined by a set of test situations (i.e., test items or test tasks) in which the behaviors can be displayed in terms of all their important nuances.

(3) given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select test tasks that will appear on any form of the test.

(4) the obtained score must be capable of expressing objectively and meaningfully the individual's performance characteristics in these classes of behavior.

Criterion-referenced tests have been used most often in instructional contexts, and in particular, in instructional procedures which seek to be individualized with respect to the learner. It is in the context of individualized instruction that questions have been raised concerning the meaning of scores on traditional educational achievement tests and their applicability to instructional decision-making. This, in turn, has led some to question the applicability of traditionally used test construction procedures. We are led to believe by some that, in individualized instruction, traditional test construction procedures are not at all applicable.

Of concern here, then, is the use of tests in adaptive individualized instructional systems. By "adaptive individualized instruction" it is meant that the instructional system is so organized and managed that the content and method of instruction varies with the individual characteristics of the student. In its ideal form, adaptive individualized instruction uses

an analysis of a student's characteristics to guide him through a course of instruction specifically tailored to him.

In the discussion that follows, it is assumed that the course of instruction a learner wishes to undertake has been designed and that it has these characteristics:

(1) the desired outcomes of instruction have been specified and translated into defined domains of tasks. The student's performance on these tasks will form the basis for inferring his attainment of the desired outcomes.

(2) a sequence of intermediate goals has been established and these goals are arranged in a prerequisite order leading to attainment of the terminal goals of instruction.

(3) various instructional procedures (methods) have been established and are available to the learner. These instructional procedures are designed for each intermediate goal and each terminal instructional goal.

The kind of tests we will consider are those that are used to make instructional decisions about individual pupils. This will leave out a number of tests designed for such purposes as overall evaluation of the course of instruction or tests designed especially for feedback to the curriculum developer concerning course improvement.

In adaptive individualized instruction three general types of decisions need to be made by the instructor and/or pupil. These decisions might be called placement, diagnosis, and attainment decisions, respectively (Glaser and Nitko, in press). One decision concerns the placement of the pupil in the instructional sequence. If the instruction is adaptive it will avoid teaching the student that of which he already has command and will

offer him new goals to learn. The information that is needed answers the question, "Where in the instructional sequence should thesstudent begin his study?" Tests built to provide this information are specific to the content and psychological structure of the particular course of instruction with which the student is faced. (Psychological structure means ordering "behaviors in a sequence of prerequisite tasks so that competence in an early task in the sequence facilitates the learning of later tasks in the sequence" (Glaser and Nitko, in press).)

As an illustration a schematic representation of a hierarchical sequence of instruction is shown in Figure 1. The lettered boxes represent instructionally relevant behaviors that are in a prerequisite order. At the bottom of the figure (below the dotted line), are behaviors that are prerequisite to the instructional sequence at hand. These behaviors are assumed to be learned prior to the student's confrontation with this sequence. The boxes in the hierarchy bear a prerequisite relationship to each other. Thus, "E" is prerequisite to "F," and "G" is prerequisite to "H." Parallel columns of boxes are considered independent of each other from the learning sequence point of view. Behaviors "I" and "J" are the terminal outcomes for this instructional sequence. Hence, "F" and "H" are both considered prerequisite to "I" and "J," but "F" and "H" are not prerequisite to each other.

Such a hierarchical specification, when it is available, provides a good "map" on which an individual student can be located before actual instruction begins. That is, one is assuming that each student needs to be located or placed at some point in this learning sequence and that a decision has <u>not</u> been made about the teaching technique that an individual is to receive in order that he may acquire the next sequential behavior.

Information provided by a placement test would result in a profile for a student such as the one shown in Figure 2. For this hypothetical student, the profile indicates that he has learned prerequisites "A," "B," "C," and "D" and intermediate goals "E," "F," and "G" well enough to proceed with instruction on behavior "H," the next behavior in the learning sequence.

An efficient test for determining location in such sequences probably would be of the branched or tailored type, particularly if the sequence was long. The nature of such tailored tests is somewhat different than those tailored tests which seek to order or locate individuals with respect to some trait, such as, general intellectual ability. In the instructional situation one can take advantage of the psychological structure of the subject matter. Thus, if an examinee was successful on items testing one objective in the sequence, this would indicate that items from earlier objectives in the sequence would be passed as well. If the hierarchy is valid, an efficient procedure is to begin testing with those items from the middle of the sequence and to branch to upper and lower points in the hierarchy depending on the examinee's score (Ferguson, 1969).

It is probable that in individiaulized instructional systems where the curriculum sequence consists of a large number of instructional objectives, for example an entire curriculum area, such neat hierarchies do not exist. Nevertheless, some sequencing of instructional objectives is possible. An example of this is shown in Figure 3. Here an elementary mathematics curriculum has been defined in terms of approximately 350 objectives. The content has been broken down into ten topics which are roughly in a prerequisite order (from top to bottom in the figure). Further, each topic has been developed over a range of complex behaviors which are also in a rough prerequisite order (from Level A through Level G in the figure). Each cell

of the grid represents several instructional objectives, and is called a unit of instruction. Usually, the objectives in a unit can be arranged in a learning sequence that leads to a few terminal goals for that unit. The inset shows (hypothetically) how a short sequence of objectives might look for one unit of instruction. In general, within a single unit, there will be prerequisites from earlier topics and lower levels.

A student that is new to this curriculum is given a two-stage placement test (Cox and Boston, 1967). The first is a broad-range test over the curriculum. The results are used to place a student at a unit in each topic or content area. The second test consists of a placement decision about the particular objectives within each unit. The broad-range test needs to be given only once at the beginning of a course of study. After completing the first unit of study, the student is given the second-stage test for the next sequential unit. Thus, he is placed at each successive unit in the curriculum.

The broad-range test actually is a battery of tests, one for each topic. Each subtest would predict for each topic, the last unit in the sequence from A to F in which the student would be successful. The student would be given instruction in the next sequential unit for that topic. Figure 4 shows a completed first-stage placement profile for a hypothetical student. Traditional item selection procedures which seek to maximize predictive validity would seem appropriate for this type of broad-range test. If the instructional sequence within a unit is hierarchical, then one could select items from the domains that define the terminal objectives of that unit, and depend on the prerequisite nature of the hierarchy to subsume the other objectives. If no such hierarchy exists, then selecting items from the domains of all objectives would seem to be required. Care should be taken, however, in

using correlational indices, since often the absolute level of attainment of unit skills is important.

Once a student is located at various points in the course or curriculum, information is required that answers the question, "What instructional alternative will best adapt to this student's individual requirements and thus maximize his attainment of the next instructionally relevant objective?" Placement in the curriculum does not specify the methods or kinds of instruction that should be used with a particular student. Tests providing this kind of information might be called "diagnostic." If there is but a single instructional method, then this is a null decision.

One general class of tests required for this type of decision comes out of aptitude-treatment-interaction research as defined by Cronbach and Snow (Cronbach, 1957; Cronbach and Snow, 1969) and suggestions for designing them are found there. It should be noted that these tests need not be criterion-referenced.

When instruction has been completed, we are interested in whether the student has learned the objectives. More often than not, a verbal statement of an instructional objective implies that an individual ought to perform quite a large domain of tasks. This is particularly true where generalization and transfer are of primary importance. The type of test which seems to provide this kind of information is a criterion-referenced test.

In constructing such tests, empirical evidence must be provided to support any contentions that the classes of test tasks from which the test constructor samples do indeed reflect the behavior or competence of interest. This means careful tryout of items and analysis of data. Domains of items need to be carefully examined and, if necessary, stratified so that representative sampling can take place. Item analysis is used both to study the

characteristics of the items and to refine them. Elimination of items on purely statistic grounds is poor practice generally in achievement test construction and becomes increasingly serious in criterion-referenced testing. The classes or domains of tasks which define a behavior are specified before a particular form of a test is developed and to screen out some items from inclusion on a particular test will change the definitions of the behavior categories (cf. Osburn, 1968). There is some evidence that tests constructed from carefully defined domains of items possess reasonably good psychometric properties without prior statistical selection (e.g., Ebel, 1962; Hively, Patterson, and Page, 1968).

Some would claim that criterion-referenced tests, particularly those that attempt to measure a single instructional objective, ought to be homogeneous. It is well known that insistence on high item total-test correlations may lower the content-validity or representativeness of the test (Cronbach, 1970a). Further, there appears to be no inherent reason why the behavior classes specified by an instructional objective need to be homogeneous (cf. Cronbach, 1970b). The opposite may be true of instructional objectives dealing with generalization and transfer of learning where one is concerned in determining proficiency on a wide variety of tasks. In most cases with these types of instructional tests, examination of item statistics, particularly in light of previous instructional history, does reveal where items or the instruction can be improved.

A further point in constructing criterion-referenced tests for measuring the outcomes of instruction is that of determining mastery. It has been mentioned that criterion-referenced testing does not necessarily imply that a cut-off score be used. These tests are used to determine the performance characteristics of the examinee with respect to the defined domain of tasks. What seems to be needed is to determine what level of

performance (or what universe score in the Cronbach sense) is required at each point in the learning sequence in order to maximize success at the next point in the sequence (Nitko, 1970). There seems to be no inherent reason why this could not differ for the individual and with the circumstances. Different students and different instructional methods may need different levels of proficiency either to continue with instruction or at the termination of instruction. The mastery score with respect to a domain of instructionally relevant tasks thus appears to be a transfer of learning problem.

Summary

In short, it is the use to which test results are put that determines their nature and the construction methodology. In instruction, various procedures cannot be considered independently of the instructional context in which they will be used. Particularly important is the integration of test design with instructional design.

References

Cox, R. and Boston M. E. Diagnosis of pupil achievement in the Individually Prescribed Instruction Project. Working Paper 15. Pittsburgh, Pa.: Learning Research and Development Center, University of Pittsburgh, 1967.

Cronbach, L. J. The two disciplines of scientific psychology. American Psychologist, 12, 1957, 671-684.

Cronbach, L. J. Essentials of Psychological Testing (3rd edition). New York: Harper and Row, 1970(a).

Cronbach, L. J. Validation of educational measures. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1970(b).

Cronbach, L. J. and Snow, R. E. Final report: Individual differences in learning ability as a function of instructional variables. Stanford University, 1969.

Ebel, R. L. Content-standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Ferguson, R. L. A model for computer-assisted criterion-referenced measurement. Education, 81, 25-31.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R. and Cox, R. C. Criterion-referenced testing for the measurement of educational outcomes. In R. Weisgerber (ed.), Instructional process and media innovation. Chicago: Rand-McNally, 1968, 545-550.

Glaser, R, and Klaus, D. J. Proficiency measurement: Assessing human performance. In R. Gagne (ed.), Psychological principles in systems development. New York: Holt, Rinehart, and Winston, 1962, 419-474.

Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (ed.), Educational Measurement. Washington, D.C.: American Council on Education. (in press)

Hively, W., Patterson, H. L., and Page, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Nitko, A. J. Criterion-referenced testing in the context of instruction. Paper presented at the Educational Records Bureau - National Council on Measurement in Education Symposium, New York, October, 1970.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

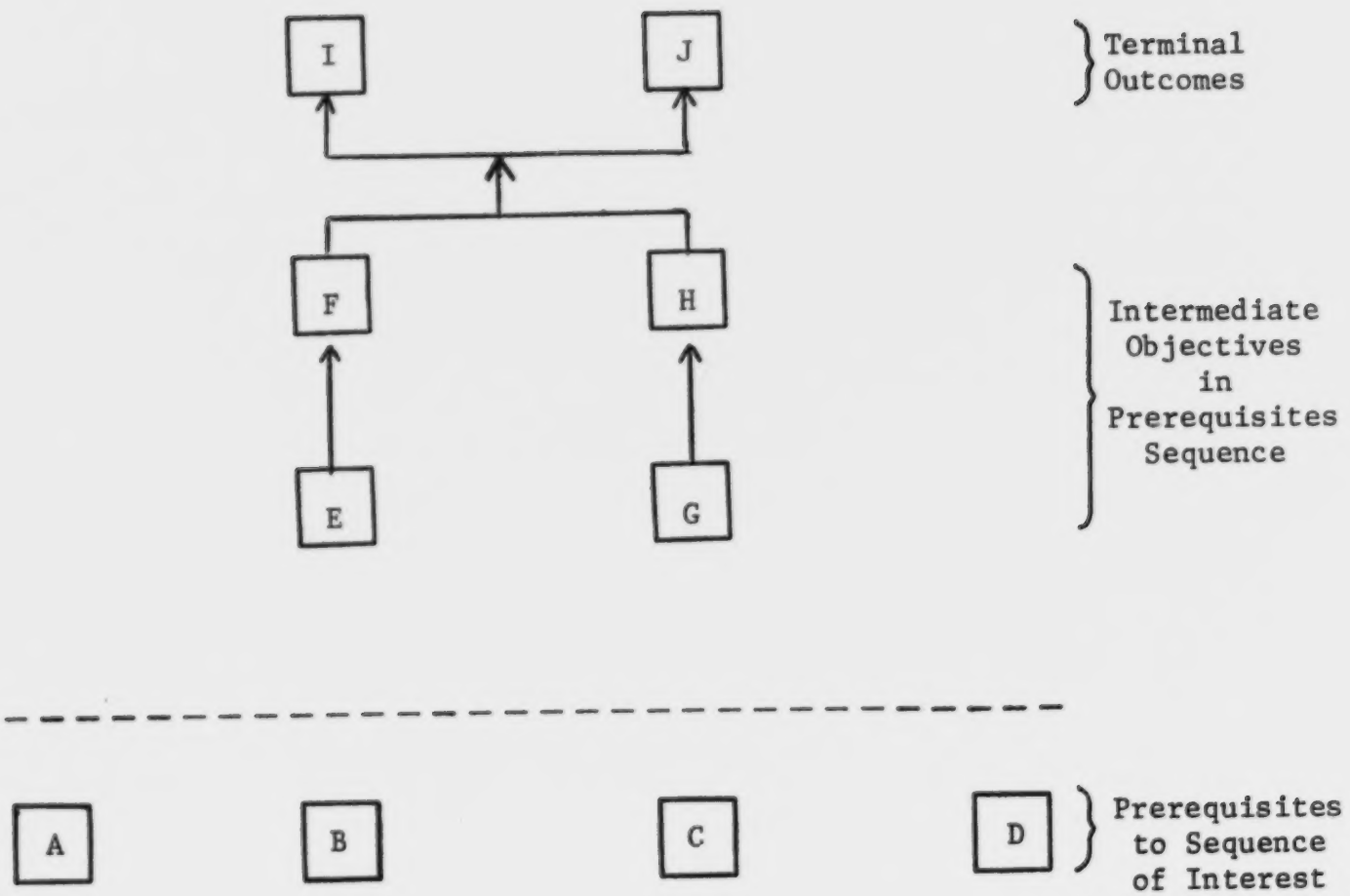


Figure 1

Hypothetical Hierarchy for a Sequence of Instruction Leading to Terminal Learning Goals "I" and "J"

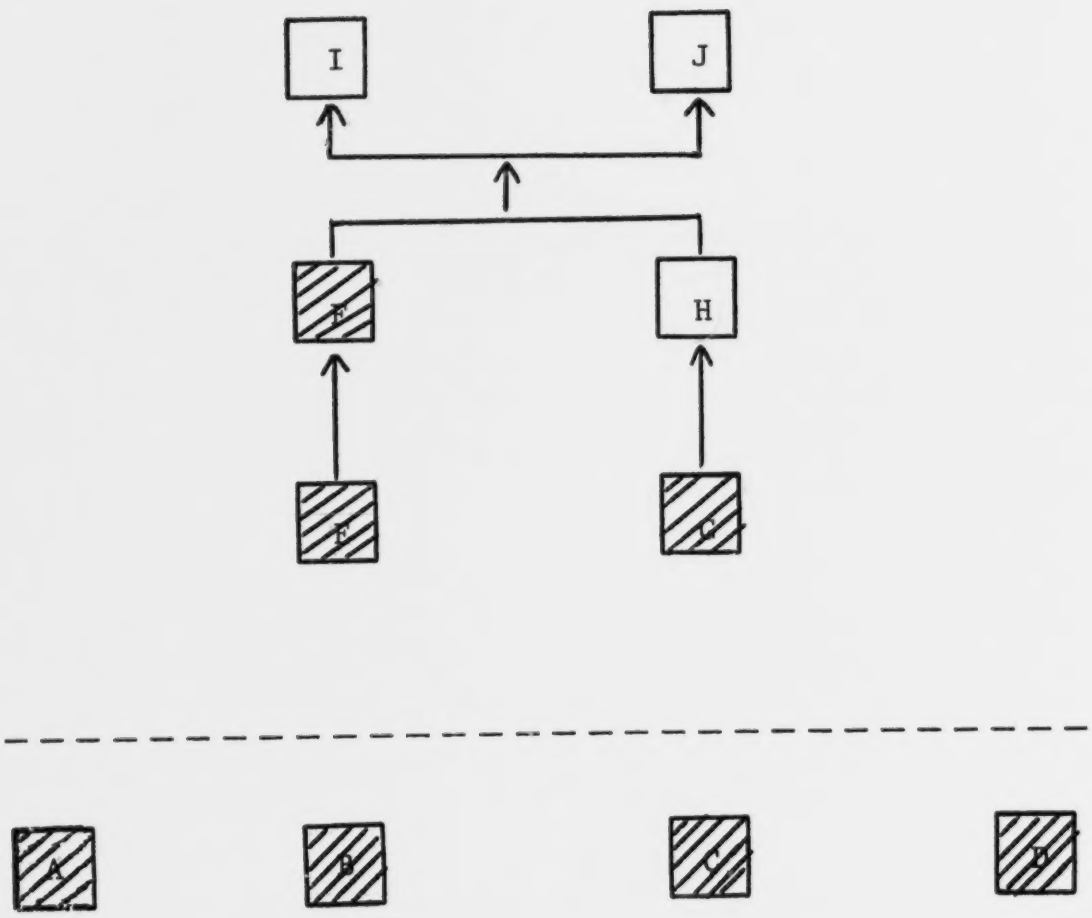
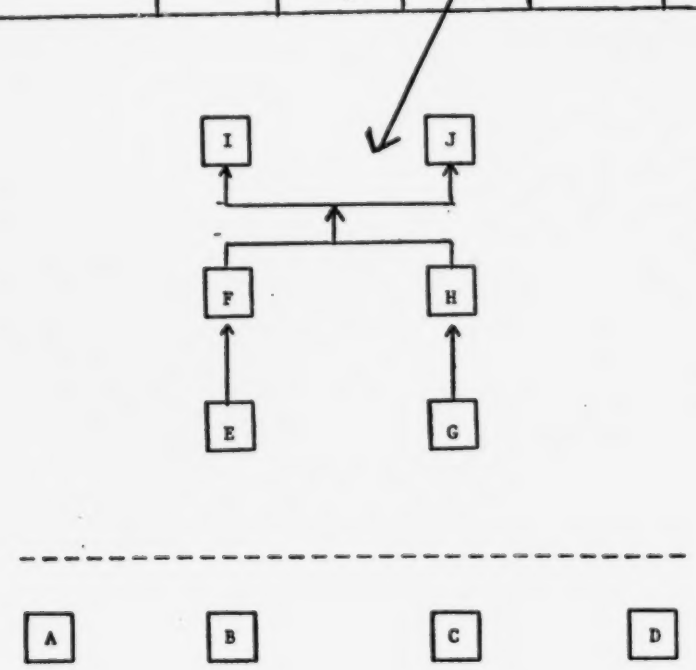


Figure 2

Placement profile for a hypothetical student. (Shaded boxes mean that the student has sufficient mastery of these instructional goals to proceed with a new instructional goal.)

Content (Topic)	Level of Complexity						
	A	B	C	D	E	F	G
Numeration/Place Value	*	*	*	*	*	*	*
Addition/Subtraction	*	*	*	*	*	*	*
Multiplication		*	*	*	*	*	*
Division		*	*	*	*	*	*
Fractions	*	*	*	*	*	*	*
Money	*	*	*	*			
Time	*	*	*	*	*		
Systems of Measurement		*	*	*	*	*	*
Geometry		*	*	*	*	*	*
Applications		*	*	*	*	*	*



*Indicates a unit of instruction consisting of one or more instructional objectives.

Figure 3

Example of curriculum layout for Individually Prescribed Instruction elementary mathematics

MATHEMATICS PLACEMENT PROFILE

Name John Smith Date 5/70 Grade 5

School Sweetdate Teacher Mrs. Jones Room 12

Mathematics Area	Placement Level A-G							Placed at Level
	A	B	C	D	E	F	G	
Numeration/Place Value								E
Addition/Subtraction								7
Multiplication								E
Division								D
Fractions								D
Money								--
Time								--
Systems of Measurement								7
Geometry								E
Applications								D

Figure 4

Example of Placement Profile for a hypothetical student with respect to the mathematics curriculum of Individually Prescribed Instruction